A Study of trends in tennis matches

Yihan Ma, Wenyu Liu, Junjie Li

Jinan University and Birmingham University Joint Institute, No.855, East Xingye Avenue, Panyu District, Guangzhou, China

Keywords: nonlinear autoregressive neural network, flow of play, model prediction, Wimbledon

Abstract: This study aims to accurately anticipate the shifts in a match by analyzing the flow of play. To achieve this, our model has been enhanced by scrutinizing the factors that impact its forecasting capabilities. To capture the flow of play, an A-value has been defined and a decision tree model has been developed. Additionally, we have built a Nonlinear Autoregressive Neural Network to fulfill the forecasting function. During the model improvement process, we calculated the Pearson correlation coefficient to gauge the extent of impact. The results indicate that the model performs its predictive function with relative success and that aces, double faults, and unforced errors are the key influencing factors.

1. Background

In today's world, people are increasingly seeking a higher quality of life and turning to tennis as a sport of choice. This demanding sport requires both physical and mental fitness, as players must react quickly to maintain an advantageous position in the game.

Momentum is a key factor in predicting a player's chances of winning a set and is influenced by a variety of factors including the serving side, score, and strength of the match. By analyzing the data available, we can gain insights into the impact of momentum on the trends and direction of tennis matches.

2. Construction of the model

Our analysis unfolds based on the 2023-wimbledon-1301 race, and we take the analysis for player 1 as an example.

2.1. Quantification of the flow of a play

2.1.1. The construction of the decision tree model

When it comes to measuring momentum, a parameter called A has been created and a decision tree model has been built to predict its value. Various factors have been examined and different scenarios have been sorted into categories, each with its unique impact on the value of A. We've defined the relative A-value as the difference between player 1 and player 2's A-values, so we can track the flow of the game based on the sign and absolute value of the relative A-value.

We've used consecutive points scored and the serving side to establish the first and second levels of decision-making, assigning the significance of A to the player's overall performance. By analyzing the A-values for both players, their performance could be assessed.

Calculating the value of A in this model follows a specific rule. Initially, the A-value is 0 and remains unchanged when there's no scoring streak. However, when there is a scoring streak, the value of A can increase by 1, 1.5, or 2 depending on whether there are two serves, one serve, or no serve. This algorithm is mapped out in Figure 1's flowchart.

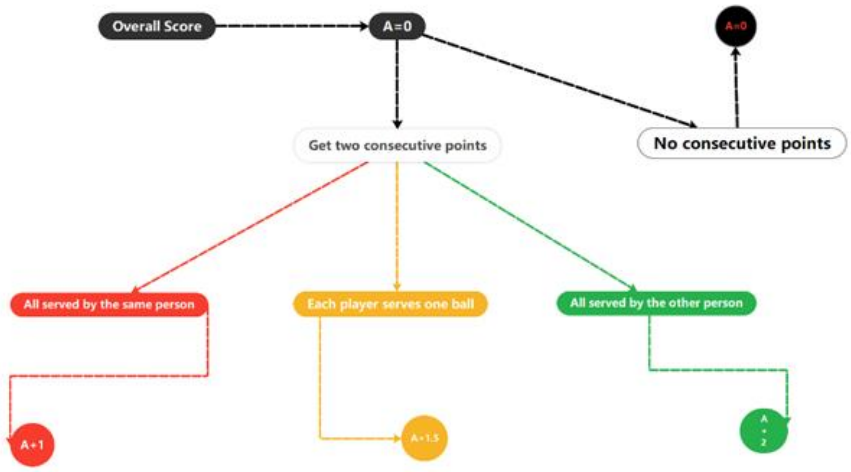
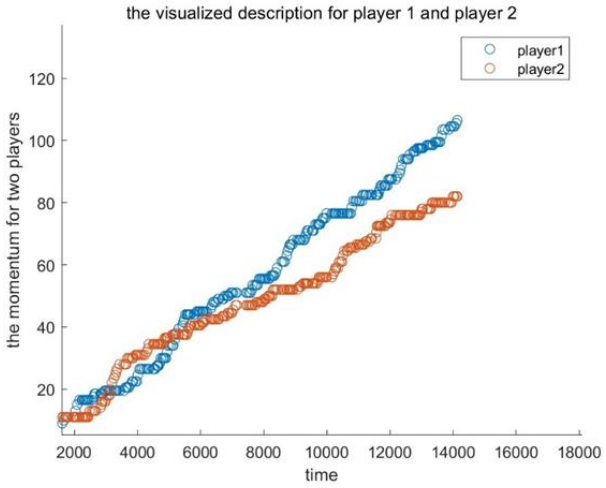


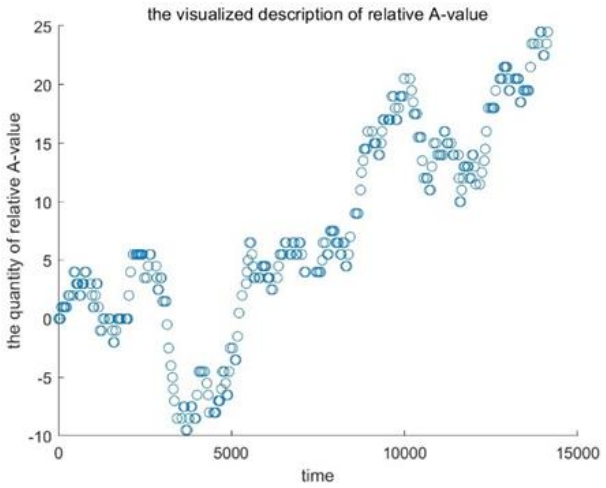
Figure1:the algorithm of the decision tree model

Subsequently, we continue to delve into the potential implications of the relative A-value. The plus and minus symbols represent the prevalence of player1 and player2, respectively, while the absolute value of the relative A-value illustrates the extent to which the dominant player outperforms the other. The accompanying visual aids, graph1 and graph2, provide a clear

depiction of the fluctuations in A-value between the two players, as well as the changes in the relative A-value.



Graph1:the visualized description for players 1 and 2



Graph2: the visualized description of relative A-value

2.2. Forecast for the swings in a game

A nonlinear autoregressive neural system has been constructed to forecast outcomes accurately. To achieve this, we have incorporated the relevant A-value of the match and divided the data into three sets: 70% for training, 15% for validation, and 15% for testing. The corresponding neural network structure is depicted in Figure 2.

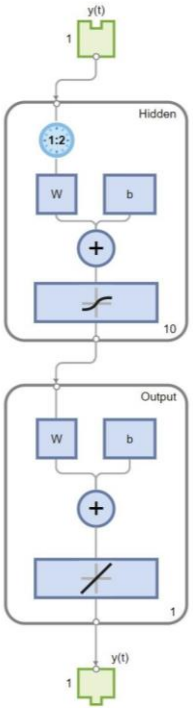
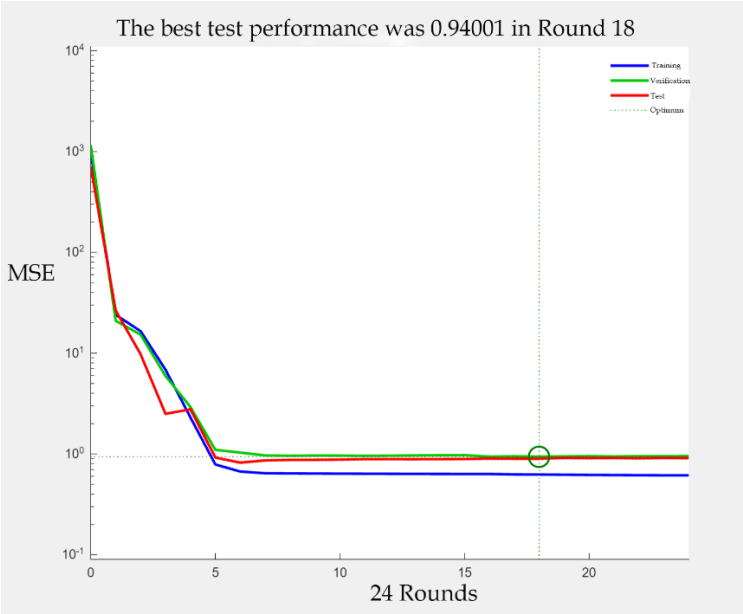


Figure2:the structure of the neural network

After final training, validation, and testing, the feedback is as follows:

	observed value	MSE	R
Training	213	0.6252	0.9958
Validation	39	0.9400	0.9936
Testing	45	0.8966	0.9936

We focus our inquiry on the MSE indicator:

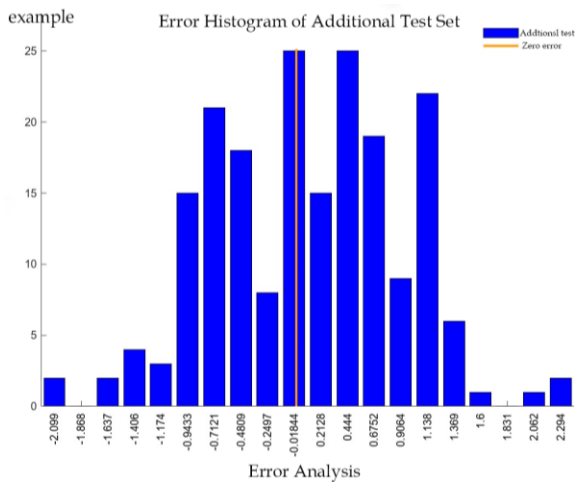


It can be seen that to some extent, the model is reliable in predicting the fluctuations in this match.

3. Tests for model generalizability

3.1. Test for other matches in the same tournament

To test the model's generalizability, we set the 2023- Wimbledon-1302 race to be the test data, and similarly, the relative A-values are calculated and imported into the neural network, and the results are as follows.

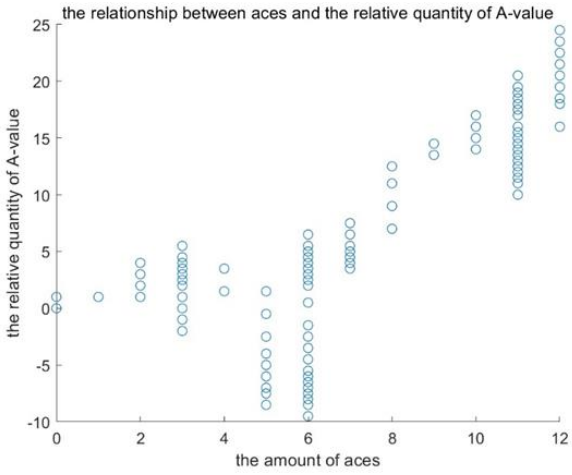


The data depicted in the figure reveals that only a handful of examples, less than five, had errors as low as 2.294 amounts. Additionally, the supplementary tests showed that over 75% of the sample sizes had minimal errors compared to the actual values, and approximately 25 samples had negligible errors. Concerning the neural network, the MSE value was 0.68, indicating that the model is capable of accurately predicting relative A-values and effectively forecasting gameplay.

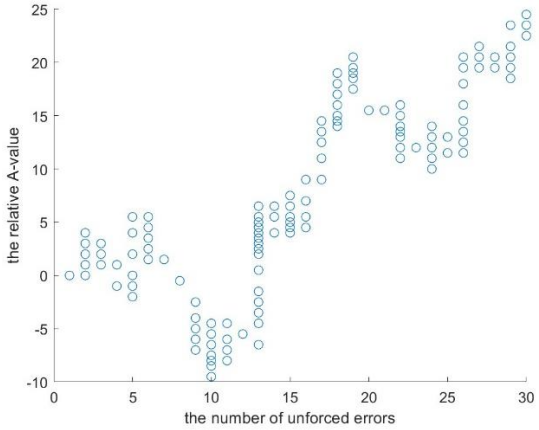
3.2. An exploration of the factors affecting model accuracy

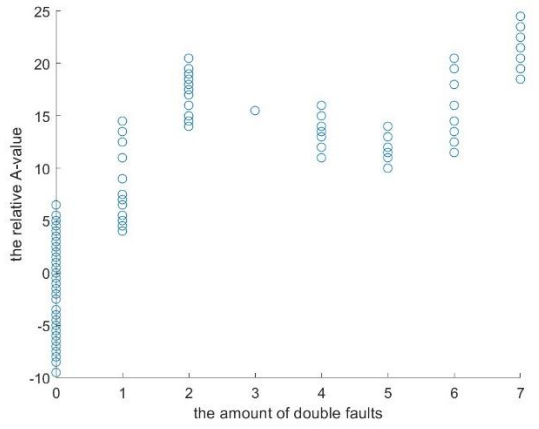
Yinman Zhang(2009) suggests that the number of aces, double faults, and unforced errors may play a pivotal role in determining the outcome of men's tennis hard court matches. In order to conduct further analysis, a Pearson correlation coefficient was utilized. To illustrate it, the number of aces accumulated after each ball was played was counted and chosen as an independent variable. Similarly, the relative A-value at each moment was determined as the

dependent variable. This allowed for a clear visualization of the function between the two variables, as shown below.



The same logic can be used to obtain the image of the function between the double faults, the number of unforced errors, and the relative A-value.





Therefore, we can calculate the Pearson correlation coefficient for each function respectively. The result shows as follows:

factor	the coefficient value
double faults	0.8123
unforced errors	0.7939
aces	0.8034

The results of the analysis revealed that the Pearson correlation coefficient indicates a high degree of association between the three variables of interest, namely double faults, unforced errors, and aces. Specifically, all three variables exhibit a correlation coefficient value of approximately 0.80, suggesting that they are the primary determinants of the outcome being studied.

4. Strengths and weaknesses of the model

- Our model is precious due to its versatility. It can be applied in various competitions and scenarios. By consolidating multiple factors into a single framework, our model allows us to accurately predict game flow and momentum.
- In certain cases, essential data may be missing, requiring us to make informed assumptions when developing our models. Accessing more comprehensive data resources would undoubtedly lead to better results. However, it is important to note that striking a balance between realism and elegance can be a challenging task, and our model prioritizes realism.

5. References

[1] COMAP Mathematics Competitions. (n.d.). Retrieved from https://www.comapmath.com/MCMICM/index.html

[2] Zhang, Yinman. (2009). Winning factors in hard court matches of the world's best men's tennis singles players. Journal of Beijing Sport University,32(10),135-137.